Evaluating probability forecasts for GDP declines using alternative methodologies

By

Kajal Lahiri[a]*, J. George Wang[b]
[a]*Department of Economics, University at Albany – SUNY, Albany, NY 12222, USA*
[b]*Department of Business, College of Staten Island – CUNY, Staten Island, NY 10314, USA*

Abstract

Evaluation methodologies for rare events from meteorology, psychology and medical diagnosis are used to examine the value of probability forecasts of real GDP declines during the current (Q0) and each of the next four quarters (Q1-Q4) using data from the *Survey of Professional Forecasters*. We study the quality of these probability forecasts in terms of calibration, resolution, odds ratio, the relative operating characteristic (*ROC*), and alternative variance decompositions. Only the shorter-term forecasts (Q0-Q2) are found to possess significant skill in terms of all measures considered, even though they are characterized by excess variability and lack of calibration.

The battery of diagnostic statistics cataloged in this paper should be useful for evaluating regression models with binary dependent variables, particularly when the event of interest is relatively uncommon.

***Key Words*:** Binary prediction, Rare events, Survey of Professional Forecasters, Subjective probability, Calibration, Resolution, Skill score, Relative Operating Characteristics, Odds ratio, Recession.

JEL Classification: B22; C11; C53

*Corresponding author: Tel: (518) 442 4758; Fax: (518) 442 4736. *E-mail addresses*: klahiri@albany.edu (K. Lahiri); george.wang@csi.cuny.edu (J. Wang)

1. Introduction

The purpose of this paper is to study the usefulness of the subjective probability forecasts that are reported in the *Survey of Professional Forecasters* (*SPF*) as predictors of GDP downturns using several distinct evaluation methodologies. Even though these forecasts are available since the fourth quarter of 1968, and have drawn academic and media attention, very little systematic analysis has been conducted in recent years to look comprehensively into their usefulness as possible business cycle indicators.[1] Typically, probability forecasts for binary outcomes are evaluated using one or two overall measures of forecast quality like the quadratic probability score (QPS) or the log probability score (LPS) over either the whole or selected spectrum of the forecasts. However, it is now well recognized that in order to examine the multifaceted nature of forecast quality, a number of alternative approaches to forecast verification should be undertaken (Murphy and Wilks 1998)). For instance, many goodness-of-fit measures can often fail to identify the ability of a forecasting system to evaluate the odds of the occurrence of a low probability event against its non-occurrence, which can be a critically important characteristic to its user.[2]

Forecasting rare or relatively uncommon business events such as recessions or financial crises has long been a challenging issue in business and economics. Since probabilistic measurement of forecasts embody important additional information over point forecasts with respect to underlying uncertainty and are increasingly being used in economic calculations, a comprehensive analysis of these forecasts can hopefully help define limits to the usefulness of macroeconomic forecasts, provide us insight into the reasons for forecast failures, and suggest pathways for possible improvement, *cf.* Granger (1996).

The plan of this paper is as follows: In section 2, we will introduce the data. In section 3, we will evaluate the probability forecasts using the traditionally popular QPS

[1] Braun and Yaniv (1992), Graham (1996) and Galbraith and van Norden (2011) reported a number of accuracy measures based on subsets of the data, and suggested that at horizons beyond two quarters these forecasts may have little skill. Lahiri and Wang (2006) analyzed the joint probability of GDP declines in the current and next quarters as a recession indicator. Clements (2008, 2009, 2011) and Clements and Harvey (2010) have studied other aspects of the data including their underlying uncertainty, internal consistency, effect of rounding, and the information superiority of aggregate over individual forecasts.

[2] See Doswell *et al.* (1990), Murphy (1991) and Stephenson (2000) for more discussions on this issue.

and calibration approach with statistical tests. In section 4, we will explore the multi-dimensional nature of the probability forecasts using alternative methodologies developed in meteorology and psychology. In section 5, we will suggest some effective ways to evaluate the performance of the probability forecasts of rare business events in terms of odds ratio and Relative Operating Curve (*ROC*) that have a long history in medical diagnosis and signal detection. Finally, concluding remarks will be summarized in section 6.

2. SPF probability forecasts of Real GDP decline

The Survey of Professional Forecasters (SPF) has been collecting subjective probability forecasts of real GDP (real GNP prior to 1992) declines during the current and four subsequent quarters since its inception in 1968:4, thanks to the foresight of Victor Zarnowitz.[3] This is about the same time that the analysis of probability forecasts got its major methodological boost when the US National Weather Service mandated the reporting of the probability of precipitation forecasts in 1965. Formerly the surveys were carried out under the auspices of the American Statistical Association and the National Bureau of Economic Research (ASA-NBER). Since June 1990, the Federal Reserve Bank of Philadelphia has conducted the survey, and maintains the data base.

The SPF respondents report their forecasts at the end of the first month of each quarter. The number of respondents has varied between 15 and 60 over the quarters. Since our aim in this study is to evaluate the macroeconomic use of SPF probability forecasts, we use forecasts averaged over individuals.[4] We also report a number of results using the primitive forecasts of two most frequent SPF respondents to supplement our analysis based on aggregate data.

[3] In addition to these probability forecasts, SPF also collects density forecasts for a number of macro variables including real GDP growth. Since from these density forecasts one can extract the implied probability of real GPD declines, Clements (2009) has studied the internal consistency of these two sets of probability forecasts at the 4-quarter horizon, and found some discrepancy.

[4] There is strong empirical evidence that combined probability forecasts result in superior predictive performance, see Graham (1996) and Clemen and Winkler (2007). Clements and Harvey (2010) have shown that at shorter horizons the average probabilities encompass practically all the useful information in the individual forecasts. The average probabilities also do not suffer from rounding as reported by Clements (2011) and Manski and Molinari (2010).

To construct the quarterly real GDP growth in real time, we used the initial or the first GDP release that is reported one month after the end of a quarter with the growth rates constructed based on lagged real GDP data available in that vintage. The target variable assumes the value 1 if the growth in that particular quarter is negative and 0 otherwise. During our sample period from 1968:4 to 2011:1, there were 24 quarters of negative GDP growth -- those beginning in 1969:4, 1970:4, 1974:1, 1978:1, 1979:2, 1980:2, 1981:2, 1982:4, 1990:4, 2001:3, and 2008:3 -- which consist of eleven separate episodes of real GDP declines. Thus, only about 14.1% in the entire sample of 170 quarters exhibited negative GDP growth based on first release of the real time data. We also conducted our analysis using five other vintages of the actual GDP data – revised one quarter, five quarters, and nine quarters after the initial release, and also revised as of March 2012. Results were virtually the same, with the initial first-release showing a very slight edge over the rest. The latter finding was also reported in Lahiri and Wang (2006). All real time data (RTDSM) were downloaded from the Federal Reserve Bank of Philadelphia web site.

In Figure 1 we have presented the average probability forecasts for real GDP declines from 1968:4 to 2011:1 for the five horizons. The shades do not represent NBER defined recessions, but the individual quarters of real GDP declines based on the first release real time data. Several interesting time series patterns are worth noting. First, the fluctuations in the probabilities over time coincided roughly with the actual GDP declines, and varied from as high as over 90% to as low as less than 5% for current quarter forecasts.[5] However, the maximum recorded probability decreased steadily from over 90% for the current quarter to only about 30% for three and four-quarter-ahead forecasts, with the forecast variance decreasing sharply from 0.057 to only 0.004 over the five horizons (see Table 1). The variations steadily got muted as the horizons increased, and by the 4-quarter horizon, hardly any of the fluctuations and lead times observed with current quarter forecasts is discernible. These observations suggest that the information content, hence the value, of the SPF probability forecasts may be horizon-dependent.

[5] Note that 1978Q1, which is the only quarter with a negative real GDP episode that was seemingly unanticipated even by the current quarter forecast, turned into a positive growth quarter in the revision reported five quarters after the initial release and remained so thereafter. Also, due to the Federal Government Shutdown, the first vintage of real GDP for 1995Q4 is missing, so we used the following quarterly vintage as an approximation.

Some additional characteristics of the data are also presented in Table 1. For instance, the average forecasted probability of a real GDP decline is higher than the average percentage of times real GDP actually fell over the whole period for all horizons - for the current quarter the overestimation is in excess of 40%.

3. Evaluation of probability forecasts

3.1 Quadratic Probability Score

A measure-oriented global approach simply compares the forecast probability with the realization of the event that is represented by a dummy variable taking value 1 or 0 depending upon the occurrence of the event. A most commonly used measure is Brier's Quadratic Probability Score (*QPS*), a probability analog of mean squared error, *i.e.*:

$$QPS = 1/T \sum_{t=1}^{T} (f_t - x_t)^2 \qquad (1)$$

where f_t is the forecast probability of the occurrence of the event at time t, x_t is the realization of the event (1 if the event occurs and 0 otherwise) at time t. T is the total number of the observations or forecasting quarters in our case. It is a *proper* score in the sense that the score is minimized only when the forecaster is truthful about his/her beliefs, and cannot benefit from hedging.[6]

The *QPS* ranges from 0 to 1 with a score of 0 corresponding to perfect accuracy, and is a function only of the difference between the assessed probabilities and realizations. The calculated *QPS* for each forecasting horizon from the current quarter (Q0) to the next four quarters (Q1, Q2, Q3, and Q4) are calculated to be 0.0668, 0.0897, 0.1065, 0.1213, and 0.1270, respectively. As expected the values deteriorate as the forecast horizon increases.

3.2 The Brier Skill Score

One way to make sense out of the raw QPS scores is to look at the QPS of a set of forecasts in relation to an unskilled baseline forecast. The most common reference

[6] Jolliffe and Stephenson (2008) have shown that Brier's score does not satisfy the property of *equitability* in the sense that all unskilled forecasters of a certain type (e.g., constant forecasts) may not yield the same expected score. The Peirce score introduced later satisfies this property, see Manzato (2007).

forecast is the unconditional probability of the event of interest known as the base rate. We calculated the bellwether skill measure defined as

$$SS(f,x) = 1 - [QPS(f,x)/QPS(\mu_x,x)] \quad (2)$$

where $QPS(\mu_x, x)$ is the mean squared error associated with the constant base rate or the constant relative frequency forecast (CRFF) μ_x, which is estimated as = 0.141 in our sample. For forecasting horizons Q0-Q4 (with CRFF QPS value of approximately 0.122), we found the SS values to be 0.45, 0.26, 0.13, 0.01, and -0.10, respectively. Note that the use of this historical average value as the base rate presumes some prior knowledge on part of the forecasters.[7] While the skill scores for the shorter run forecasts (Q0-Q2) indicate substantial improvement over the benchmark base rate forecast, the longer run forecasts (Q3 and Q4) do not show any clear-cut relative advantage.

Lopez (2001) generalized the equal forecast accuracy test (S_1) of Diebold and Mariano (1995) to probability forecasts. Using this principle we can test the null hypothesis that for any particular horizon, the QPS of CRFF is the same as the QPS of SPF forecasts, i.e., the skill score is zero. If $\bar{d}$ represents the average of

$$d_t = 2(x_t - f_t)^2 - 2(x_t - 0.141)^2 \quad (3)$$

and $f_d(0)$ is the spectral density function at frequency zero, then asymptotically $S_1 = \frac{\bar{d}}{\sqrt{2\pi f_d(0)/T}}$ is $\mathcal{N} \sim (0,1)$ under the null. In recent years many authors have emphasized the importance of forecast serial correlation on the sampling properties of Brier score or the Brier skill score, see Pesaran and Timmermann (2009), Wilks (2010) and Blöchlinger and Leippold (2011). Ironically, the incidence of the problem gets worse when forecasts are more skillful. The advantage of the Diebold-Mariano test is that a variety of features of forecast errors including non-zero means, non-normality, contemporaneous correlation, and more importantly a serially correlated QPS differential are readily accommodated. Using *Stata* version 12.1 (with Maxlag=13 chosen by Schwert criterion, and uniform kernel), we calculated the S_1 statistics to be: -2.564 (p-value 0.01), -1.942 (p-value 0.05), -

[7] Alternatively, one can consider using the last realization as the forecast to form a binary time varying base rate. Thus, for current and next four quarters, the last quarter realization is used. The associated skill scores of SPF forecasts were significantly less than those with CRFF of 0.141 implying that the latter base rate is considerably more informative than the use of the lagged actual. Other base rate alternatives, *e.g.*, eternal optimist (*f*=0), eternal pessimist (*f*=1), or a coin flipper (*f*=0.5), are also considerably less informative than the CRFF alternative; *cf.* Zellner *et al.* (1991).

1.540 (p-value 0.12), -0.269 (p-value 0.78), and 0.652 (p-value 0.51) for Q0, Q1, Q2, Q3, and Q4 respectively. Thus, the current and one quarter-ahead forecasts have statistically significant skill over the constant base rate forecast, whereas the Q3 and Q4 forecasts clearly show no statistical evidence of any skill over CREF. The quarter 2 forecasts are seen to be only marginally skillful with a p-value of 0.12.

We also studied the forecast skill of two SPF forecasters (ID#65 and ID#84) who are the top two most frequent participants in the survey spanning almost the whole period. Beginning in 1968Q4, forecaster #65 responded in 106 quarters before dropping out of reporting these probabilities in 2006Q4. Forecaster #84 participated in 121 surveys before dropping out in 2009Q4. Their skill scores respectively were (.009, - .087, -0.135 -.317 and -.447) and (.409, .220, .183, -.118 and -.207) for current and subsequent four quarters. Surprisingly, even with such a long period of survey participation, forecaster #65 exhibits little skill even for the current quarter. The overall pattern for Forecaster # 84 is very similar to that of the average probabilities where the S_1 statistics for first three horizons were found to be significant at the 10% level of significance. Note that the two forecasters did not respond over the same quarters, hence their scores are not directly comparable.

3.3 Prequential Test for Calibration

The QPS is a global measure that cannot discern differential skills over different probability intervals. Dawid (1984) and Seillier-Moiseiwitsch and Dawid (1993) (henceforward *SM-D*) suggested a test for calibration-in-the-small (i.e., over a set of mutually exclusive probability ranges) when a sequence of *T* probability forecasts are grouped in probability intervals and are compared with observed relative frequencies for each group. Let us denote $r_j = \sum_{t=1}^{T_j} x_{tj}$ ($j = 1, \ldots J$) as the actual number of times the event actually occurred when the number of issued probabilities belonging to the probability group j (with midpoint m_j) is T_j ($T = \sum_{j=1}^{J} T_j$), and x_{tj} is the binary outcome index (=1 if the event occurs; =0, otherwise) for the t^{th} occasion in which forecast f_j was offered. Then $e_j = m_j / T_j$ is the expected number of occurrences in the group j. The *SM-D* test

statistic for calibration or accuracy in each probability group is obtained as the weighted difference between the predictive probability and the realization of the event $Z_j = (r_j - e_j)/\sqrt{w_j}$, where w_j is the weight determined by $w_j = T_j m_j (1-m_j)$. If the Z_j statistic lies too far out in the tail of the standard normal distribution, it might be regarded as evidence against forecast accuracy for the probability group *j*. The overall performance of the forecasts for all *j* can be evaluated using the test statistic $\sum Z_j^2$, which is distributed asymptotically χ^2 with *j-1* degrees of freedom. Thus, a forecaster would exhibit good calibration-in-the-small if on 70% of the times when he or she forecasts a 0.7 chance of a GDP decline, it actually declines, if 10% of the times when he or she forecasts a 0.1 chance of a GDP decline, it actually declines, and so forth. Using the *SM-D* calibration test, the accuracy of probability forecasts can be statistically assessed with explicitly expressed uncertainty as indicated by the confidence level. It should be noted that the *SM-D* test, under reasonably weak conditions, is asymptotically independent of the process that generated the forecasts including certain serial dependence. This is an important point because quarters of negative real GDP growth typically tend to cluster, and hence may not be stochastically independent.[8]

The detailed calculations using current quarter forecasts with eleven probability intervals are presented in Table 2. We observe that only a few probability intervals show significant miscalibration, and many of the probability intervals contain very few observations. At longer horizons Q1-Q4, an increasing number of the higher probability intervals become empty. Following *SM-D*'s guideline to ensure adequate number of observations in each range, we consolidated the eleven probability bins into only three groups and computed the calibration test statistics for each forecast horizon. These calculations are reported in Table 3. The calculated χ^2 values are such that forecasts for all quarters are found to be not calibrated at the 5% level of significance with 2 degrees freedom (critical value: 5.99). Unfortunately, due to the sparseness of observations at higher probability bins, we cannot make any inference on calibration at different deciles above 0.25 forecasted probabilities. Galbraith and van Norden (2011) report a similar

[8] Apparently being unaware of the SM-D test, Blöchlinger and Leippold (2011) have developed a similar test of calibration, and successfully applied to the problem of credit default forecasting.

result on global miscalibration using kernel-based non-parametric methods. Note that the non-parametric approach does not circumvent the problem of sparseness of high probability forecasts. Braun and Yaniv (1992) do not report statistical tests, but find evidence that the historical base rate forecast is better calibrated than SPF forecasts at all available horizons. Braun and Yaniv (1992) experimented with a number of other conditional real time base-rate models, but the results were qualitatively the same as with the simple CREF base rate.

Note that the observed lack of calibration cannot necessarily be attributed to aggregation of probabilities over forecasters, cf. Ranjan and Gneiting (2010). We calculated the *SM-D* χ^2 test statistic for forecasters ID#65 and ID#84 for all five horizons. Again, due to lack of recorded forecasts in many of the bins, we had to consolidate them into three bins (0.0-0.149, 0.15-0.249 (0.15 -0.349 for Q0), and 0.25-1.00 (0.35-1.00 for Q0)). The χ^2 values ranged from 10.56 to 52.98 for forecaster # 65 and from 16.32 to 73.10 for forecaster # 85 for the five horizons respectively, which, on being referred to a χ^2 with 2 degrees of freedom, fall decidedly in the rejection region. Thus, like the aggregate forecasts, these primitive forecasts of the two most frequent respondents were also found to be miscalibrated at all horizons.

The overwhelming reason for the observed miscalibration is that over majority of these probability intervals, forecasters tend to assign more probability for a GDP decline than the observed historical proportions would warrant. This is consistent with the fact that the average probability forecast over the whole sample is higher than the proportion of negative growth quarters in the sample; for instance μ_x for Q0 is 0.141 against μ_f = 0.199 (see Table 1). Considering the fact that recessions are mostly under anticipated, this empirical finding may seem surprising, *cf.* Fildes and Stekler (2002). But note that most of the times our forecasters issue non-zero (*albeit* small) probabilities for negative GDP growth even in periods of positive growth (see Table 4d). Since nearly 85% of our sample quarters are characterized by positive real GDP growth, the overestimation of the probability of a negative GDP growth during these quarters overwhelms the full sample

result. In the psychological literature on calibration, overconfidence has been the dominant finding, and a number of behavioral explanations have been offered.[9]

4. Further diagnostic verifications

4.1 The Murphy Decomposition

In addition to calibration, there are several other features that also characterize the quality of probability forecasts. Murphy (1972) decomposed the *QPS* or the Brier Score into three components:

$$QPS(f,x) = \sigma_x^2 + E(\mu_{x|f} - f)^2 - E(\mu_{x|f} - \mu_x)^2 \qquad (4)$$

where $E(.)$ is the expectation operator, σ_x^2 is the variance of x_t, and $\mu_{x|f}$ is the conditional mean of x given the probability forecast f. Rewriting *QPS* for grouped data

$$QPS(f,x) = (1/T)\sum_{j=1}^{J}\sum_{t=1}^{T_j}(f_j - x_{tj})^2, \qquad (5)$$

which acts like a crude form of non-parametric smoothing, the Murphy decomposition in (5) can be expressed as

$$QPS(f,x) = \bar{x}(1-\bar{x}) + (1/T)\sum_{j=1}^{J} T_j(\bar{x}_j - m_j)^2 - (1/T)\sum_{j=1}^{J} T_j(\bar{x}_j - \bar{x})^2 \qquad (6)$$

where $\bar{x}_j = (1/T_j)\sum_{t=1}^{T_j} x_{tj}$ is the relative frequency of event's occurrence over T_j occasions

with forecast m_j, *i.e.*, $\bar{x}_j$ (= r_j/T_j) is an estimate of $\mu_{x|f}$ using grouped data.

The first term on the RHS of (6) is the variance of the observations, and can be interpreted as the *QPS* of constant forecasts equal to the base rate. It represents forecast difficulty. The second term on the RHS of (6), introduced already in section 3.3, represents the calibration or reliability of the forecasts. For a person to be perfectly calibrated, assessed probability should equal percentage of correct forecasts over a number of assessments. The third term on the RHS of (6) is a measure of resolution or

[9] See, for instance, Braun and Yaniv (1992), Ungemach et al. (2009), Erev et al. (1994) and Wright and Ayton (1992).

discrimination; it refers to the ability of a set of probability forecasts to sort individual outcomes into probability groups which differ from the long-run relative frequency. It refers to an aspect of forecasting skill that discriminates individual occasions on which the event of interest will and will not take place. Note that a sample of probability forecasts will be completely resolved only if the forecast probabilities only take values zero and one. For perfectly calibrated forecasts, i.e., $\mu_{x|f} = m$ and $\mu_x = \mu_f$, the resolution term equals the variance of the forecasts, σ_f^2. Forecasts possess positive *absolute* skill when the resolution reward exceeds the miscalibration penalty in (6). Even though calibration is a natural feature to have, it is resolution that makes the forecasts useful in practice, *cf.* DeGroot and Fienberg (1983).[10]

Numerical values of the Murphy decomposition are given in Table 4a where we find that *QPS* improves by about 41%, 23% and 12% for the current (Q0), one quarter-ahead (Q1), and 2-quarter–ahead (Q2) forecasts, respectively, over the constant relative frequency forecast (CRFF). The 3-quarter-ahead (Q3) forecasts are almost even with CRFF, and the *QPS* of the 4-quarter-ahead (Q4) forecasts is slightly worse than the Q3 value, as expected. The major contributor for the improvement in *QPS* is resolution, which helps to reduce the baseline *QPS* (CRFF) by about 51%, 26%, 22%, 4%, and 1% for Q0 to Q4, respectively. On the other hand, the miscalibration increases *QPS* of CRFF by 10%, 4%, 10%, 3% and 4%, respectively – they are relatively small and almost the same for all forecast horizons. Braun and Yaniv (1992) reported a similar result using data up to 1988:Q3. Thus, even though *SM-D* tests rejects calibration, this rejection does not substantially deteriorate the QPS scores. It is resolution that dominates QPS of the probability forecasts.

Note that given (1) and (2), the Brier Skill Score in (4) can be written as resolution minus calibration - both terms scaled by σ_x^2. Thus, the scaled resolution directly measures the fraction of σ_x^2 that would be explained by the forecasts in the absence of calibration

[10] Dawid (1986) presents a simple example to distinguish between calibration and resolution. Suppose the event (=1) occurs in every alternative period as 0, 1, 0, 1, …. Consider three sets of forecasts: F_1 assigns 0.5 always; F_2 assigns 0, 1, 0, 1, …, and F_3 assigns 1, 0, 1, 0,…. Here F_1 and F_2 are well calibrated, but F_2 is perfect whereas F_1 is almost useless. Both F_2 and F_3 are perfectly resolved, but F_3 is not well calibrated. F_3 is more useful than F_1 once we know how to calibrate F_3.

error. These numbers, presented in Table 4a in the parentheses under resolution column for the five horizons respectively, tell a similar story.

Murphy (1988) suggested another useful decomposition of the skill score defined in (2):

$$SS(f,x) = \rho_{fx}^2 - [\rho_{fx} - (\sigma_f / \sigma_x)]^2 - [(\mu_f - \mu_x)/\sigma_x]^2 \qquad (7)$$

where ρ_{fx} is the correlation coefficient between forecast and the actual binary outcome, σ_f^2 and σ_x^2 are their true variances, and (μ_f, μ_x) are the respective means, see also Murphy and Winkler (1992). The decomposition shows that SS is simply the square of the correlation between *f* and *x* adjusted for any miscalibration penalty (second term) and the normalized difference in the sample averages of the actual and the forecast (third term). This decomposition for Q0-Q4 with ungrouped data are given in Table 4b where we find that the last two terms of the decomposition are close to zero, and thus, the skills for Q0-Q4 forecasts in effect reflect the correlations between the forecasts and the actual binary outcomes. For Q0, Q1 and Q2 these correlations are 0.477, 0.294 and 0.150 respectively, and are found to be statistically significant using the simple t-test for no correlation at the 5% level of significance.[11] The correlation for Q3 is statistically insignificant, and that for Q4 was small, negative and significant. This decomposition again shows that the long-term probability forecasts have no statistical skill compared to the benchmark forecast, and it is resolution and not calibration that is important for the observed skill of Q0 - Q2 forecasts.

4.2 The Yates Decomposition

The calibration component in (6) can be written as:

$$(1/T)\sum_{j=1}^{J} T_j (m_j - \bar{x}_j)^2 = s_f^2 + (\bar{f} - \bar{x})^2 - 2s_{fx} + (1/T)\sum_{j=1}^{J} T_j (\bar{x}_j - \bar{x})^2 \qquad (8)$$

where $\bar{f}$, s_f^2 and s_{fx} are the sample forecast mean, variance and covariance respectively. Since the last term in equation (8) is the resolution component in (6), Yates (1982) and

[11] The appropriate t-values were obtained from a regression of (ρ_{fx}/σ_x^2) *x* on *f* augmented by one lagged values of the dependent and the independent variables to accommodate serial autocorrelation, see Pesaran and Timmermann (2009).

Yates and Curley (1985) have suggested a covariance decomposition of *QPS* that is more basic than the Murphy decomposition, see also Yates (1994). The so-called Yates decomposition is written as:

$$QPS(f,x) = \mu_x(1-\mu_x) + \Delta\sigma_f^2 + \sigma_{f,\min}^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x} \qquad (9)$$

where $\sigma_{f,\min}^2 = (\mu_{f|x=1} - \mu_{f|x=0})^2 \mu_x(1-\mu_x)$, and $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$.

As noted before, the outcome index variance $\sigma_x^2 = \mu_x(1-\mu_x)$ provides a benchmark reference for the interpretation of *QPS*. The conditional minimum forecast variance $\sigma_{f,\min}^2$ is the minimum of σ_x^2 that is necessary to support a given wedge between $\mu_{f|x=1}$ and $\mu_{f|x=0}$, and reflects the double role that the variance of the forecast plays in forecasting performance. Even though minimization of σ_f^2 will reduce *QPS*, this minimum value of forecast variance will be achieved only when a constant forecast is offered. But a constant forecast would lead to zero covariance between the forecast and event, which will increase *QPS*. Note that $\sigma_{f,x}$ can be expressed as $(\mu_{f|x=1} - \mu_{f|x=0})\mu_x(1-\mu_x)$, and constant forecast generates $(\mu_{f|x=1} - \mu_{f|x=0}) = 0$. So, for a given $(\mu_{f|x=1} - \mu_{f|x=0})$ or $\sigma_{f,x}$, the solution is to minimize the forecast variance, which demonstrates the fundamental forecast ability of the forecasters. The value of the conditional minimum forecast variance $\sigma_{f,\min}^2$ becomes σ_f^2 when $\Delta\sigma_f^2$ is zero, i.e., when the forecaster has perfect foresight such that (s)he can exhibit perfect discrimination of the instances in which the event does and does not occur.

Since $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$, this term is considered as the excess variability in forecasts. If the covariance indicates how responsive the forecaster is to information related to an event's occurrence, $\Delta\sigma_f^2$ might reasonably be taken as a reflection of how responsive the forecaster is to information that is not related to the event's occurrence. Note that $\Delta\sigma_f^2$ can be expressed as $(T_1\sigma_{f|x=1}^2 + T_0\sigma_{f|x=0}^2)/T$, where $T_i (i = 0,1)$ is the number of periods associated with the occurrence (i =1) and non-occurrence (i = 0), $T_1 + T_0 = T$. So the term is the weighted mean of the conditional forecast variances.

Using the SPF probability forecasts, the components of equation (9) were computed and presented in Table 4c. Note that the *QPS* values and the variances for Q0-Q4 in Tables 4a and 4c are slightly different because the Yates decomposition could be done with ungrouped data whereas the Murphy decomposition was done with probabilities grouped as in Table 2.[12] For the shorter forecasting horizons up to 3-quarters (Q0-Q3), the overall *QPS* values are less than the constant relative frequency forecast variance, which demonstrate the absolute skillfulness of the SPF probability forecasts. For the 4-quarter forecasting horizon (Q4), the overall *QPS* is slightly higher than that of the constant relative frequency forecast. The primary contributor of the performance is the covariance term that helps reduce the forecast variance by almost 94%, 52%, 24% and 6% for up to 3-quarter-ahead forecasts, but makes no contribution for the 4-quarter-ahead forecasts. The covariance reflects the forecaster's ability to make a distinction between individual occasions in which the event might or might not occur. It assesses the sensitivity of the forecaster to specific cues that are indicative of what will happen in the future. It also shows whether the responsiveness to the cue is oriented in the proper direction. This decomposition is another way of reaching the same conclusion as the decomposition of skill score in Table 4b.

The excess variability of the forecasts, $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$, for each horizon is found to be 0.0298, 0.0204, 0.0104, 0.0044, and 0.0037, respectively. Compared to the overall forecast variances 0.0567, 0.0287, 0.0122, 0.0045, and 0.0037, the excess variability's of SPF probability forecasts are 53%, 71%, 85%, 98% and 100% for Q0-Q4 forecasts, respectively. Thus, they are very high, and this means that the subjective probabilities are scattered unnecessarily around $\mu_{f|x=1}$ and $\mu_{f|x=0}$, see Figure 2. Since the difference in conditional means, $\mu_{f|x=1} - \mu_{f|x=0}$, are very close to zero for Q3-Q4 forecasts, almost all of their variability is attributed to excess variability. Assigning low probabilities in periods when GDP actually fell seems to be one of the root causes of the excess variance. In our sample real GDP declined twenty four times. However, in sixteen of these

[12] Note that due to the grouping of probabilities in Murphy decomposition, two within-bin variance and covariance components should be added to the resolution term calculated in the Murphy decomposition to make the measure independent of binning, see Stephenson et al. (2008). By comparing the QPS scores in Tables 4a and 4c, we however find that the joint effect of these two additional terms is very small, possibly because nearly 90% of the responses are rounded to multiples of 5 or 10, see Clements (2011).

quarters, the assigned probabilities for Q1 forecasts never exceeded 0.5; for Q3-Q4 forecasts, the assigned probabilities were even below 0.2 in most of these instances. In contrast, for Q0-Q2, in more than 90% of the quarters when GDP growth did not decline, the probabilities were assigned correctly below 50% (for Q3-Q4 the probabilities were below 30%). These issued probabilities could have been even lower for better resolution. The Yates decomposition gives this critical diagnostic information about the SPF probability forecasts that the Murphy decomposition could not.

The problem with excess variability can be clearly seen if we examine the two conditional density probabilities or likelihoods given $x = 1$ (GDP decline) and $x = 0$ (no GDP decline) in Figure 2. See also Table 4d. For these two conditional distributions, the mean probabilities were calculated to be (0.60, 0.42, 0.29, 0.20 and 0.18) for $x = 1$ and (0.13, 0.16, 0.17, 0.17 and 0.18) for $x = 0$, respectively for different horizons. Also, *Var (f|x=1)* is much larger than *Var (f|x=0)* for the first three horizons where the forecasts have reasonable discriminatory power.

These empirical conditional likelihoods can be contrasted with a notional "ideal" pattern that can be generated using a simple model. Let us assume that the variable *z* generates the probability of the binary event *x*. Suppose the conditional density of *z* given *x*, denoted by $h(z|x, a, b, c)$, is *Normal* with mean $ax + c(1 - x)$ and standard deviation b. Then it can be shown (see Hastie et al. 2001, their eq. (4.9), p. 108) that x will follow the logit model,

$$P \equiv P(x = 1|z) = \Phi(\beta_0 + \beta_1 z) \qquad (10)$$

where Φ is the CDF of the standard logistic distribution, with β_0 and β_1 satisfying

$$\beta_0 = ln\frac{P(x=1)}{P(x=0)} - \frac{1}{2b^2}(a^2 - c^2), \qquad (11)$$

$$\beta_1 = \frac{1}{b^2}(a - c), \qquad (12)$$

and $P(x = 1)$ and $P(x = 0)$ are the unconditional probabilities. Then the conditional density of P given x can be written as:

$$f(P|x, a, c, b) = \frac{1}{|\beta_1|} h\left(\frac{\Phi^{-1}(P) - \beta_0}{\beta_1}\middle| x, a, c, b\right)\frac{d\Phi^{-1}}{dP} \qquad (13)$$

where Φ^{-1} is the inverse function of Φ with $P \in (0,1)$ as the argument, *cf.* Hogg and Craig (1994, p. 169). These two densities, conditional on $x = 1$ and 0 are depicted in Figure 3 for parameter values $a = -c = 1.4$ and b = 2 with $P(x = 1) = 0.5$. As we see,

they are highly skewed in the desirable directions, with most of the probability masses piled up over very low probabilities for x = 0 and over very high probabilities for x = 1. However, there is a substantial amount of overlap in the densities, crossing each other at the expected value P=0.5. By Monte Carlo integration we computed the mean values of the densities to be 0.675 for x=1 and 0.329 for x=0. By varying the values of *a* and *b* in (13) one can easily find that the degree of overlap between them, and hence the discriminatory power of the forecasts will be determined not only by the values of *a* and *c* (location) but also by the value of *b* (variance).[13] In general, good discriminatory forecasts will give two largely non-overlapping conditional distributions, and the ratio of their vertical differences should be as large as possible.

In our case (Figure 2), for the shorter run forecasts (Q0-Q2), even though the conditional densities during positive growth periods (x=0) display an almost ideal shape, the same for x=1 is highly dispersed and lacks enough mass at higher probabilities. Simply put, SPF forecasters do not issue high enough probabilities for GDP declines in periods when real GDP actually declined, even at very short horizons. The longer run forecasts (Q3-Q4) display poor discrimination due to an almost exclusive use of relatively low probabilities during periods of GDP declines (*i.e.*, $x = 1$). At long horizons, these conditional densities are degenerating into the conditional densities for x=0, and the two distributions fully overlap. In particular, the means and variances for $x = 1$ (GDP decline) and $x = 0$ (no GDP decline) for the 4-quarter ahead forecasts (Q4) are almost identical, suggesting SPF forecasters do not have the necessary information to identify forthcoming negative GDP growth periods.

Overall, both the Murphy and Yates decompositions suggest ways of improving the forecasts, particularly at short-run horizons. One obvious solution is to find ways to reduce unnecessary variance of forecasts keeping the covariance the same during both regimes, especially during GDP declines. For a given covariance, a reduction of excess variability will increase the resolution further. This is an important economic insight regarding business cycle forecasting that can be gleaned from SPF probability forecasts.[14]

[13] Cramer (1999) suggested the use of the difference in the conditional means as a goodness-of-fit measure in binary choice models.

[14] Clements (2008) finds little evidence that the asymmetry in forecasters' loss functions can possibly explain the relatively low forecast probabilities for impending real GDP declines. Rudebusch and Williams

5. Receiver operating characteristic (*ROC*)

In evaluating rare event probabilities, the impact of correctly identifying the frequent event, which often is a primary source of hedging, should be minimized. In such situations, a better approach is to concentrate on the hit rate and the false alarm rate for the infrequent event, instead of looking at "percentage correctly predicted" that is the very basis of *QPS*, *cf.* Doswell *et al* (1990) and Murphy (1991). The hit rate (*H)* is the proportion of times an event was forecast when it occurred, and the false alarm rate (*F)* is the proportion of times the event was forecast when it did not occur.

A simple and often-used measure of forecast skill, the Peirce (or sometimes called Kuipers) skill score (PS) first introduced by Peirce (1884), is obtained by taking the difference between the hit rate (*H*) and the false alarm rate (*F*). Based on a set of continuous probability forecasts, the decision to issue a forecast for the occurrence or non-occurrence of an event is typically made based on a predetermined threshold (say, *w*) on the weight of evidence scale *W*. The occurrence forecast is announced if $W > w$, the non-occurrence is announced otherwise. Given the decision threshold *w*, a 2x2 contingency table for successes and failures for the classifier can be constructed to organize the information, as shown in Table 5. Here *a* represents the number of quarters of forecasted GDP declines that actually happened, *b* shows forecasted but not observed, *c* is for observed quarters of GDP declines that were not forecasted, and *d* is the correct forecast of the non-event when real GDP did not fall. Varying the threshold value *w* changes these four statistics, but the system has only two degrees of freedom, which are often characterized by H and F. Then the PS can be calculated, for a given threshold *w*, as $\text{PS}(w) = H - F = (ad - bc)/((a+c)(b+d))$. Assuming independence of the hit and false alarm rates, the asymptotic standard error of PS is given by

---

(2009) show that the SPF forecasters could improve their Q3 and Q4 forecasts by incorporating readily available information in the yield curve.

$\sqrt{(H(1-H)/(a+c))+F(1-F)/(b+d)}$; see Agresti (2007). Alternatively, based on the market-timing test of Pesaran and Timmermann (1992), Granger and Pesaran (2000) have suggested an asymptotic N (0, 1) test for the significance of the Peirce test, $PT=\sqrt{T}PS/\sqrt{P_x(1-P_x)/\bar{x}(1-\bar{x})}$, where $P_x=\bar{x}H+(1-\bar{x})F$.

Stephenson (2000) has argued that the forecast skill for rare events can be judged better by comparing the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm), *i.e.*, by using the odds ratio $\theta = \{H/(1\text{-}H)\}/\{F/(1\text{-}F)\}$ which is simply equal to the cross-product ratio $(ad)/(bc)$ obtainable from the contingency table. The odds ratio is unity when the forecasts and the realizations are independent or PS=0, and can be easily tested for significance by considering the log odds that is approximately *Normal* with a standard error given by $\sqrt{1/a+1/b+1/c+1/d}$. Note that each cell count should be at least 5 for the validity of the approximation. PS and θ are reported in Table 6 for relevant values of the decision threshold.

One important but often overlooked issue in the evaluation of probability forecasts for uncommon events is the role of the selected threshold (w) – the binary event classifier. The performance of a probability forecast of such events in terms of discrimination ability is actually the result of the combination of the intrinsic discrimination ability of a forecasting system and the selection of the prevalence-dependent threshold. In these regards, Receiver Operating Characteristic (*ROC*) is a convenient descriptive approach that can be very informative in many situations.[15]

ROC can be represented by a graph of the hit rate against the false alarm rate as w varies, with the false alarm rate plotted as the X-axis and the hit rate as the Y-axis. The location of the entire curve in the unit square is determined by the intrinsic discrimination capacity of the forecasts, and the location of specific points on a curve is determined by the decision threshold w that is selected by the user. As the decision threshold w varies from low to high, or the *ROC* curve moves from right to left, H and F vary together to

[15] This approach, developed during World War II to assess radar receivers in signal detection, has a long history in medical imaging and in evaluating loan default and rating forecasts, *cf.* Hanley and McNeil (1982) and Stein (2005). See Jolliffe and Stephenson (2003) for additional analysis on the use of ROC, and Berge and Jordà (2011) for a recent business cycle application. Fawcett (2006) raises a number of issues in the use of AUC – area under ROC – as a standalone measure of discrimination across all possible ranges of thresholds.

trace out the *ROC* curve. Low thresholds lead to both high *H* and *F* towards the upper right hand corner. Conversely, high thresholds make the *ROC* points move towards the lower left hand corner along the curve. Thus, a perfect discrimination is represented by an *ROC* that rises from (0, 0) along the *Y*-axis to (0,1), then straight right to (1,1). The diagonal $H = F$ bisector represents zero skill, indicating that the forecasts are completely non-discriminatory. *ROC* points below this no-skill line represent the same level of skill as they would if they were located above the diagonal, but are just mislabeled, *i.e.*, a forecast of non-occurrence should be taken as occurrence.

In Figure 4 the *ROC* curves together with their 95% confidence bands for the current quarter and the next four quarters are displayed. These were calculated using *rocfit* on *Stata*, see Ma and Hall (1993) and Argesti and Coull (1998). [16] It can be seen that the *ROC* for the current quarter (Q0) is located maximally away from the diagonal towards the left upper corner demonstrating the highest discrimination ability of the SPF forecasts, followed by the one-quarter-ahead forecasts. For longer-term forecasts ROCs become rapidly flatter as the forecasting horizon increases. For the four-quarter-ahead forecasts, the ROC curve wraps around the diagonal line staying very close over the whole range, and the associated confidence band suggests that it is statistically not distinguishable from the diagonal line. This means that the 4Q forecasts have no skill or discrimination ability for any value of the threshold.

In situations where the analyst may have only a vague idea about the relative costs of type I and type II errors (*e.g.*, in the problem of predicting the turning point in a business cycle), (s)he can pick a comfortable hit rate (or false alarm rate), and the underlying ROC curve will give the corresponding false alarm rate (or hit rate). This will also give an optimal threshold for making decisions. When the relative costs two types of errors are known exactly, the decision theoretic framework developed by Zellner *et al*. (1991), Granger and Pesaran (2000) and others can be used to issue recession forecasts.[17] However, before using the probability forecasts in decision-making, the significance of their skillfulness should first be established.

[16] See Hall et al. (2004) for obtaining nonparametric confidence intervals. We are currently looking into the possible effect of temporal dependence on these intervals.

[17] Elliott and Timmermann, (2008) present a systematic treatment of forecast value in a decision theoretic framework. See Roebber and Bosart (2006) for two weather-related examples comparing forecast value against forecast skill.

The hit rates and false alarm rates for selected threshold values in the range 0.50-0.05 are reported in Table 6, where one can find the mix of hit and false alarm rates that are expected to be associated with each horizon-specific set of forecasts issued by SPF.[18] For example, for achieving a hit rate of 92% with Q0 forecasts, one can use 0.25 as the threshold, and the corresponding false alarm rate is expected to be 0.15. Table 6 also shows that at this threshold value, even though the false alarm rates are roughly around 0.15 for forecast of all horizons, the hit rate steadily declines from 92% for Q0 to only 13% for Q4 - clearly documenting the rapid speed of deterioration in forecast quality as the forecast horizon increases. Though not explicitly reported in Table 6, for the same hit rate of 92% across all horizons, the false alarm rates for Q0 through Q4 forecasts were found to be 0.221 (*w=0.19)*, 0.236 (*w=0.20)*, 0.555 (*w=0.14)*, 0.769 (*w=0.12*) and 0.913 (w=0.10) respectively. Thus, for the same hit rate of 92%, the corresponding false alarm rates for Q3-Q4 forecasts are so large (77% and 91% respectively) that they can be considered useless for all practical purposes, and thus, may have very little value in decision-making.

In Table 6 we have also reported the Peirce scores (PS) and the odds ratios (θ) for selected *w*. The rapid decline in these values as the forecast horizon increases is remarkable, and for Q4 forecasts these values are close to zero and unity respectively, suggesting no-skill. Note that for *w* values between 0.20 and 0.30, PS values (and also odds ratios) are maximized for all horizons. Manzato (2007) has shown that, under certain conditions, the PS-maximizing threshold is the probability value at which the two conditional likelihood functions given $x = 0$ and $x = 1$ intersect making the likelihood ratio equal to one, provided they are unimodal and have some significant overlap. This expectation is borne out reasonably well in likelihood diagrams in Figure 2, except for the approximation errors due to the finite conditional samples and the histograms we are working with. Note that the use of the conventional threshold $w = 0.50$ would have yielded much lower hit rates (0.67 for Q0; 0.33 for Q1; 0.04 for Q2; 0.00 for Q3 and Q4), but with a very low false alarm rate of around 0-4 percent. As is expected, this is indeed the threshold that maximized the percentage of correct classifications for all five

[18] In order to save space, we did not report in Table 6 the values of *w* greater than 0.5. Moreover, these values were not relevant in our context.

horizons, and can be considered inappropriate under different loss functions, cf. Cramer (1999).

Using the critical value 1.645 for a one-sided normal test at the 5% level, the PS and θ values were found to be statistically significant for Q0-Q2 and insignificant for Q4 forecasts.[19] Based on the standard error formula $\sqrt{[H(1-H)/(a+c)]+[F(1-F)/(b+d)]}$ for PS reported in Agresti (2007), PS values for Q3 were insignificant at the 5% level for all allowable values of *w*. Same was the conclusion using the odds ratio test with *w* = 0.15 for which OR value was the maximum. However, the PT test for Q3 was statistically significant for *w* = 0.25 at the 5% level. Note that these tests do not correct for the presence of clusters in the binary data, and recent research has established that tests that do not correct for such serial dependence will generally be oversized and tend to over-reject the null, see Pesaran and Timmermann (2009) and Wilks (2010). Following Pesaran and Timmermann (2009), we followed their dynamically augmented reduced rank approach by regressing *x* on *f* augmented by lags of *x* and *f* as additional explanatory variables. By AIC model selection criterion, only one lagged *x* and f were needed in the augmented regression where the residuals became serially uncorrelated. These tests indicated that Q0-Q2 forecasts remained highly significant, but now Q3 joins Q4 as being statistically not significant at the 5% level.[20] Thus, the weight of our evidence suggests that Q3 and Q4 forecasts have practically no statistically significant skill. We should emphasize that statistical significance or insignificance does not mean the forecasts have utility or value in a particular decision theoretic context.[21] In the current context, however, the use of Peirce skill as the utility function identifies a prevalence-based threshold point on the ROC curve that maximizes the skill of the classifier, and yields reasonable hit and false alarm rates for the near term forecasts.

In Table 6, we have also presented the relevant parts of the ROC statistics and associated skill scores for forecasters with IDs 65 and 84. The tradeoffs between hit rates

[19] The significance tests were conducted only for cases where the cell counts were more than 5.

[20] The test results were exactly the same when we estimated the standard errors by a simple bootstrap procedure, and also where the observations were sampled independently stratified by the binary dependent variable. The bootstrap replications were set at 1000 and the bootstrap sample size was the same as the original sample size. We used *Stata* to do these calculations.

[21] Granger and Pesaran (2000) show how, under certain simplifying assumptions, PS can be used as an indicator of economic value.

and false alarm rates, the Peirce scores and odd ratios are very similar to those associated with the average probabilities. However, the forecaster # 84 is somewhat better than the forecaster # 65 when we compare their PS or hit rates for the same false alarm rates. For Q2, forecaster # 84 looks very similar to the average forecaster in that both exhibit some positive forecasting skill, and perform much better than forecaster # 65. Another important finding is that, like the average probability, these two individual forecasters also do not have any forecasting skill beyond Q2.

6. Conclusions

We have evaluated the subjective probability forecasts for real GDP declines during 1968:4-2011:1 using alternative methodologies developed in such diverse fields as psychology, meteorology, and medical diagnosis. Thus the terrain is necessarily vast, and we touched only on the major promontories directly relevant for our analysis. Each methodological approach is shown to have contributed differently to our understanding of the nature of these forecasts.

We decomposed the traditional *QPS* score into calibration, resolution, and alternative variance decompositions. We found overwhelming evidence that the shorter run forecasts at horizons Q0-Q2 possess significant skill having good discrimination ability, but are not well calibrated. Across probability bins, the assigned probabilities on the average were significantly higher than conditional relative frequencies for each of the five forecast horizons. Interestingly the lack of calibration did not affect the overall forecasting performance as measured by the QPS score in an important way; when forecasts have value it is due to good resolution that describes how far away the forecasts are from their average relative frequency. However, we found that the variances of the forecasts, compared to their discrimination capability, are significantly more than necessary, particularly during cyclical downturns. The analysis of probability forecasts, thus, shows that forecasters respond *also* to cues that are not related to the occurrence of negative GDP growths.

Across all skill measures, the Q3 forecasts are found to have borderline value, if at all. The Q4 forecasts exhibit poor performance as characterized by negative skill scores, low resolutions, dismal ROC measures, and insignificant correlations with actual

outcomes. It is clear from our analysis that the professional forecasters do not possess adequate information to forecast meaningfully at horizons beyond two quarters; they lack relevant discriminatory cues. Since the SPF panel is composed of professional economists and business analysts who forecast on the basis of models and informed heuristics, their failure for the long-term forecasts may indicate that at the present time forecasting declines in real GDP beyond two quarters may not be possible with reasonable type I and type II errors. Isiklar and Lahiri (2007) reached a very similar conclusion using point forecasts for real GDP growth from multi-country Consensus Economics survey data.

We have emphasized that in evaluating rare event probabilities like recessions, the impact of correctly identifying the frequent event, which often is a primary source of hedging, should be minimized. In such situations, a better approach is to concentrate on hit and false alarm rates of the infrequent event, instead of the statistic "percentage correctly predicted" that is the very basis of global measures like *QPS.* In this context, we found the ROC curves to be useful in our context, where the relative odds for the event can be studied at depth, and thereby an end user's loss function for missed signals and false alarms can be considered in making decisions. The ROC analysis in our case revealed that for a pre-assigned hit rate of (say) 90%, the associated false alarm rates for the Q3-Q4 forecasts for negative GDP declines are so high (in excess of 80%) that they may be considered useless for all practical purposes.

ACKNOWLEDGMENTS:

An earlier version of this paper was presented at the First ifo/INSEE/ISAE Macroeconomic Conference, Rome, and at the Far Eastern Meeting of the Econometric Society, Tsinghua University, Beijing. We are grateful to three referees, the associate editor, editor Graham Elliott, Roy Batchelor, Michael Clemens, Richard Cohen, Nigel Harvey, David Stephenson, Herman Stekler and Kenneth Wallis for many helpful comments and suggestions. Yongchen Zhao and Yang Liu provided generous research assistance. We are solely responsible for any errors and omissions.

References:

Agresti, A. (2007). *An introduction to categorical analysis,* Second Edition, John Wiley and Sons: New Jersey.

Agresti, A., & Coull, B.A. (1998). Approximation is better than "exact" for interval estimation of binomial proportions. *American Statistician*, 52, 1-7.

Berge, T., & Jordà, Ò. (2011). Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics,* 3, 246-277.

Blöchlinger, A., & Leippold, M. (2011). A new goodness-of-fit test for event forecasting and it's application to credit defaults. *Management Science,* 57, 487-505.

Braun, P, & Yaniv, I. (1992). A case study of expert judgment: economists' probabilities versus base rate model forecasts. *Journal of Behavioral Decision Making*, 5, 217-231.

Clemen, R. T., & Winkler, R.L. (2007). Aggregating probability distributions. In E. Ward, R. F. Miles, & D. von Winterfeldt (eds), *Advances in Decision Analysis: from Foundations to Applications* (pp.154-176), Cambridge: Cambridge University Press.

Clements, M. P. (2008). Consensus and uncertainty: Using forecast probabilities of output declines. *International Journal of Forecasting*, 24, 76-86. .

Clements, M. P. (2009). Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In J. L. Castle, & N. Shephard (eds), *The Methodology and Practice of Econometrics*. A Festschrift in Honour of David F. Hendry. (Chapter 8, pp. 206 – 226). Oxford University Press.

Clements, M. P. (2011). An empirical investigation of the effects of rounding on the SPF probabilities of decline and output growth histograms. *Journal of Money, Credit and Banking*, 43, 207-220.

Clements, M. P., & Harvey, D. I. (2010). Forecasting encompassing tests and probability forecasts. *Journal of Applied Econometrics,* 25, 1028-1062.

Cramer, J. (1999). Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society,* Series D (*The Statistician*), 48, 85-94.

Dawid, A. P. (1984). Statistical Theory: A prequential approach. *Journal of the Royal Statistical Society,* 147, 279-297.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson & C. B. Mead (eds.), *Encyclopedia of Statistical Sciences* (Vol. 7, pp.210-218), New York: Wiley-Interscience.

DeGroot, M. H., & Fienberg, S. E. (1983). A comparison and evaluation of forecasters. *The Statistician,* 32, 12-22.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics,*13 (6), 253-263.

Diebold, F. X., & Rudebusch, G. D. (1991). Turning point prediction with the composite leading index: An *ex ante* analysis", in K. Lahiri & G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Approaches*. Cambridge University Press, Cambridge.

Doswell, C. A., Davies-Jones, R., & Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting,* 5, 576-585.

Elliott, G. & Timmermann, A. (2008). Economic forecasting, *Journal of Economic Literature,* 46(1), 3-56.

Erev, I., Wallsten, T. S., & Budescu, D. (1994). Simultaneous over- and under confidence: The role of error in judgment process. *Psychological Review,* 101, 519-527.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27, 861-874.

Fildes, R. & Stekler, H.O. (2002). The state of macroeconomic forecasting, *Journal of Macroeconomics*, 24 (4), 435-468.

Galbraith, J. W., & van Norden, S. (2011). Kernel-based calibration diagnostics for recession and inflation probability forecasts. *International Journal of Forecasting,* doi:10.1016/j.ijforecast.2010.07.004.

Graham, H. R. (1996). Is a group of forecasters better than one? than none? *Journal of Business,* 69 (2), 193-232.

Granger, C. W. J. (1996). Can we improve the perceived quality of economic forecasts? *Journal of Applied Econometrics,* 11, 455-473.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting,* 19, 537-560.

Hall, P., Hyndman, R.J., & Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrica,* 91 (3), 743-750.

Hanley, J. A., & McNeil, B. (1982). The Meaning and use of the area under the receiver operating characteristics (ROC) curve. *Diagnostic Imaging,* 29, 307-335.

Hastie, T., Tibshirani, R., & Friedman, J. (2001), *The elements of statistical learning: data mining, inference, and prediction*, Springer

Hogg, R., & Craig,A. (1994), *Introduction to mathematical statistics*, Prentice Hall.

Isiklar, G., & Lahiri, K. (2007). How far ahead can we forecast? Evidence from cross-country surveys, *International Journal of Forecasting*, 23, 2007, 167-187.

Jolliffe, I. T., & Stephenson, D. B. (2008). Proper scores for probability forecasts can never be equitable. *Monthly Weather Review*, 136, 1505-1510.

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecasting verification: A practitioner's guide in atmospheric science*, Eds. John Wiley & Son Limited.

Lahiri, K., & Wang, J. G. (1994). Predicting cyclical turning points with leading index in a Markov switching model. *Journal of Forecasting*, 13 (3), 245-263.

Lahiri, K., & Wang, J. G. (2006). Subjective probability forecasts for recessions: Guidelines for use. *Business Economics*, 41 (2), 26 - 37.

Lopez, J. A. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, 20, 87-109.

Ma, G., & Hall, W. J. (1993). Confidence bands for the receiver operating characteristics curves. *Medical Decision Making*, 13, 191-197.

Manski, C., & Molinari, F. (2010). Rounding of probabilistic expectations in surveys. *Journal of Business and Economic Statistics*, 28, 219-231.

Manzato, A. (2007). A Note on the maximum Peirce score. *Weather and Forecasting*, 22, 1148-1154.

Murphy, A. (1972). Scalar and vector partitions of the probability score: Part I. Two-state Situation. *Journal of Applied Meteorology*, 11, 273-282.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116, 2417-2424.

Murphy, A. H. (1991). Probabilities, odds, and forecasters of rare events. *Weather and Forecasting*, 6, 302-306.

Murphy, A. H., & Wilks, D. S. (1998). A Case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Weather and Forecasting*, 13, 795– 810.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7, 435-435.

Peirce, C.S. (1884). The numerical measure of the success of predictions, *Science*, 4, 453-454.

Pesaran, M. H., & Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association,* 104, 325-337.

Pesaran, M. H., & Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics,* 10, 461-465.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society*, Series B, 72, Part 1, 71-91.

Roebber, P. J., & Bosart, L. F. (2006). The complex relationship between forecast skill and forecast value: A real-world analysis. *Weather and Forecasting,* 11, 544-558.

Rudebusch, G. D., & Williams, J. C. (2009). Forecasting recessions: The puzzle of the enduring power of the yield curve. *Journal of Business and Economic Statistics,* 27, 492-503.

Seillier-Moiseiwitsch, F., & Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association,* 88 (421), 355-359.

Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking and Finance,* 29, 1213-1236.

Stephenson, D. B. (2000). Use of the 'odds ratio' for diagnosing forecast skill. *Weather and Forecasting*, 15, 221-232.

Stephenson, D. B., Coelho, C. A. S., & Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, 23, 752-757.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20, 473-479.

Wilks, D. S. (2010). Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society,* Part B, 136, 2109-2118.

Wright ,G., & Ayton, P. (1992). Judgmental probability forecasting in the intermediate and medium term. *Organizational Behavior and Human Decision, Processes* 51, 344-363.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30, 132-156.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (eds.), *Subjective Probability* (pp.381- 410), John Wiley: Chichester, UK.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analysis of probabilistic forecasts. *Journal of Forecasting,* 4, 61-73.

Zellner, A., Hong, C., & Min, C-K. (1991). Forecasting turning points in international growth rates using Bayesian exponentially weighted autoregression, time varying parameter, and pooling techniques. *Journal of Econometrics,* 49, 275-304.

Table 1: Summary Measures of Forecasts & Realizations, 1968Q4-2011Q1

Quarter	***Means***		***Variances***		***Correlation Coefficient***	***Sample Size***
	μ_f	μ_x	$Var(f)$	$Var(x)$		
Q0	*0.1998*	*0.1412*	*0.0567*	*0.1220*	*0.6907*	*170*
Q1	*0.1986*	*0.1420*	*0.0287*	*0.1226*	*0.5419*	*169*
Q2	*0.1855*	*0.1429*	*0.0122*	*0.1232*	*0.3874*	*168*
Q3	*0.1755*	*0.1437*	*0.0045*	*0.1238*	*0.1541*	*167*
Q4	*0.1764*	*0.1420*	*0.0037*	*0.1226*	*-0.1750*	*162*

Table 2: Calculations for Calibration Test: Quarter 0

Prob. Interval	Midpoint	Freq.	Occurrence	Relative Freq.	Expectation	Weight	*N* (0,1)	χ^2
	m_j	T_j	r_j	r_j/T_j	$e_j = m_j T_j$	$w_j = m_j T_j (1 - m_j)$	$Z_j = (r_j - e_j)/\sqrt{w_j}$	Z_j^2
0.00 - 0.049	0.025	48	1	0.02	1.20	1.17	-0.18	0.03
0.05 - 0.149	0.100	63	0	0.00	6.30	5.67	-2.65	7.00
0.15 - 0.249	0.200	15	1	0.07	3.00	2.40	-1.29	1.67
0.25 - 0.349	0.300	14	5	0.36	4.20	2.94	0.47	0.22
0.35 - 0.449	0.400	4	0	0.00	1.60	0.96	-1.63	2.67
0.45 - 0.549	0.500	5	2	0.40	2.50	1.25	-0.45	0.20
0.55 - 0.649	0.600	6	4	0.67	3.60	1.44	0.33	0.11
0.65 - 0.749	0.700	4	2	0.50	2.80	0.84	-0.87	0.76
0.75 - 0.849	0.800	4	2	0.50	3.20	0.64	-1.50	2.25
0.85 - 0.949	0.900	7	7	1.00	6.30	0.63	0.88	0.78
0.95 - 1.000	0.975	0	0	0.00	0.00	0.00	0.00	0.00
		170	**24**					

Table 3: Calibration Tests, Q0-Q4

Prob. Interval	Midpoint	Freq.	Occurrence	Relative Freq.	Expectation	Weight	Test Statistic
	m_j	T_j	r_j	r_j/T_j	$e_j = m_j T_j$	$w_j = m_j T_j (1 - m_j)$	$Z_j = (r_j - e_j)/\sqrt{w_j}$
Q0							
0.00 - 0.149	*0.063*	*111*	*1*	*0.01*	*7.50*	*6.84*	*-2.49*
0.15 - 0.349	*0.250*	*29*	*6*	*0.21*	*7.20*	*5.34*	*-0.52*
0.35 - 1.000	*0.696*	*30*	*17*	*0.57*	*20.00*	*5.76*	*-1.25*
Total		*170*	*24*				***Overall χ^2 = 8.01***
Q1							
0.00 - 0.149	*0.063*	*96*	*1*	*0.01*	*6.00*	*5.63*	*-2.11*
0.15 - 0.249	*0.200*	*29*	*4*	*0.14*	*5.80*	*4.64*	*-0.84*
0.25 - 1.000	*0.647*	*44*	*19*	*0.43*	*28.47*	*10.05*	*-2.99*
Total		*169*	*24*				***Overall χ^2 = 14.06***
Q2							
0.00 - 0.149	*0.063*	*79*	*3*	*0.04*	*4.977*	*4.66*	*-0.92*
0.15 - 0.249	*0.200*	*56*	*6*	*0.11*	*11.20*	*8.96*	*-1.74*
0.25 - 1.000	*0.647*	*33*	*15*	*0.45*	*12.3*	*7.54*	*0.98*
Total		*168*	*24*				***Overall χ^2 = 9.21***
Q3							
0.00 - 0.149	*0.063*	*62*	*5*	*0.08*	*3.91*	*3.66*	*0.57*
0.15 - 0.249	*0.200*	*84*	*14*	*0.17*	*16.80*	*13.44*	*-0.76*
0.25 - 1.000	*0.647*	*21*	*5*	*0.24*	*13.59*	*4.80*	*-3.92*
Total		*167*	*24*				***Overall χ^2 = 16.28***
Q4							
0.00 - 0.149	*0.063*	*63*	*9*	*0.14*	*3.97*	*3.72*	*2.61*
0.15 - 0.249	*0.200*	*78*	*11*	*0.14*	*15.60*	*12.48*	*-1.30*
0.25 - 1.000	*0.647*	*21*	*3*	*0.14*	*13.59*	*4.80*	*-4.83*
Total		*162*	*23*				***Overall χ^2 = 31.87***

Table 4a: Murphy Decomposition

Quarter	*MSE (Accuracy)*	*= Uncertainty*	*+ Reliability*		*- Resolution*	
Q0	*0.0708*	*0.1220*	*0.0132*	*(0.1085)*	*0.0644*	*(0.5282)*
Q1	*0.0931*	*0.1226*	*0.0055*	*(0.0447)*	*0.0349*	*(0.2848)*
Q2	*0.1090*	*0.1232*	*0.0109*	*(0.0884)*	*0.0250*	*(0.2032)*
Q3	*0.1228*	*0.1238*	*0.0045*	*(0.0366)*	*0.0056*	*(0.0450)*
Q4	*0.1269*	*0.1226*	*0.0057*	*(0.0462)*	*0.0013*	*(0.0107)*

Note: Numbers in the parenthesis are reliability and resolution relative to uncertainty.

Table 4b: Decomposition of Skill Score

Quarter	*SS (Skill core)*	*= Association*	*- Calibration*	*- Bias*
Q0	*0.4488*	*0.4771*	*0.0001*	*0.0282*
Q1	*0.2642*	*0.2936*	*0.0033*	*0.0262*
Q2	*0.1301*	*0.1501*	*0.0053*	*0.0148*
Q3	*0.0142*	*0.0238*	*0.0014*	*0.0082*
Q4	*-0.1002*	*0.0306*	*0.1211*	*0.0097*

Table 4c: Yates Decomposition

Quarter	$QPS =$	$VAR(x) +$	$\Delta VAR(f) +$	$MinVAR(f) +$	$(\mu_f - \mu_x)^2 -$	$2 * COVAR(f, x)$
Q0	*0.0668*	*0.1212*	*0.0298*	*0.0269*	*0.0034*	*0.1145*
Q1	*0.0897*	*0.1218*	*0.0204*	*0.0084*	*0.0032*	*0.0641*
Q2	*0.1066*	*0.1224*	*0.0104*	*0.0018*	*0.0018*	*0.0299*
Q3	*0.1213*	*0.1231*	*0.0044*	*0.0001*	*0.0010*	*0.0073*
Q4	*0.1271*	*0.1218*	*0.0037*	*0.0000*	*0.0012*	*-0.0004*

Table 4d: Summary Measures of Conditional Distributions Given Realizations

Quarter	*Means*		*Variances*		*Sample*	*Sample*
	$\mu_{f\|x=0}$	$\mu_{f\|x=1}$	$Var(f) \mid x = 0$	$Var(f) \mid x = 1$	$T_0(x = 0)$	$T_1(x = 1)$
Q0	*0.1333*	*0.6044*	*0.0237*	*0.0686*	*146*	*24*
Q1	*0.1614*	*0.4238*	*0.0181*	*0.0350*	*145*	*24*
Q2	*0.1681*	*0.2901*	*0.0097*	*0.0152*	*144*	*24*
Q3	*0.1713*	*0.2008*	*0.0046*	*0.0035*	*143*	*24*
Q4	*0.1766*	*0.1750*	*0.0038*	*0.0032*	*139*	*23*

Table 5: Schematic Contingency Table

	Event Observed		
Event Forecasted	Occurred	Not Occurred	Total
Yes	a (hit)	b (false alarm)	a+b
No	c (miss)	d (correct rejection)	c+d
Total	a+c	b+d	a+b+c+d = T

Table 6: Measures of Forecast Skill: Quarter 0 to Quarter 4

	Q0				*Q1*				*Q2*				*Q3*				*Q4*			
w	*H*	*F*	*PS*	*OR*	*H*	*F*	*PS*	*OR*	*H*	*F*	*PS*	*OR*	*H*	*F*	*PS*	*OR*	*H*	*F*	*PS*	*OR*
0.05	*0.96*	*0.68*	*0.28*	*10.92*	*1.00*	*0.92*	*0.08*	*0.00*	*1.00*	*0.99*	*0.01*	*0.00*	*1.00*	*0.99*	*0.01*	*0.00*	*1.00*	*0.99*	*0.01*	*0.00*
0.10	*0.96*	*0.40*	*0.55*	*33.92*	*0.96*	*0.60*	*0.36*	*15.33*	*1.00*	*0.78*	*0.22*	*0.00*	*1.00*	*0.87*	*0.13*	*0.00*	*0.91*	*0.91*	*0.00*	*0.00*
0.15	*0.96*	*0.25*	*0.71*	*70.28*	*0.96*	*0.34*	*0.61*	*43.70*	*0.88*	*0.47*	*0.40*	*7.82*	*0.79*	*0.60*	*0.19*	*2.52*	*0.61*	*0.61*	*0.00*	*0.99*
0.20	*0.92*	*0.20*	*0.72*	*44.38*	*0.96*	*0.23*	*0.72*	*75.09*	*0.75*	*0.22*	*0.53*	*10.94*	*0.50*	*0.27*	*0.23*	*2.67*	*0.30*	*0.35*	*-0.04*	*0.83*
0.25	*0.92*	*0.15*	*0.77*	*62.00*	*0.79*	*0.17*	*0.62*	*18.24*	*0.63*	*0.13*	*0.50*	*11.67*	*0.21*	*0.11*	*0.10*	*2.09*	*0.13*	*0.13*	*0.00*	*1.01*
0.30	*0.88*	*0.11*	*0.77*	*56.88*	*0.71*	*0.12*	*0.58*	*17.13*	*0.42*	*0.08*	*0.33*	*7.86*	*0.04*	*0.06*	*-0.02*	*0.65*	*0.00*	*0.04*	*-0.04*	*0.00*
0.35	*0.71*	*0.09*	*0.62*	*24.85*	*0.54*	*0.09*	*0.45*	*12.00*	*0.21*	*0.06*	*0.15*	*3.95*	*0.00*	*0.03*	*-0.03*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
0.40	*0.71*	*0.07*	*0.64*	*33.03*	*0.50*	*0.07*	*0.43*	*13.50*	*0.17*	*0.05*	*0.12*	*3.91*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
0.45	*0.71*	*0.06*	*0.65*	*36.97*	*0.42*	*0.06*	*0.36*	*12.23*	*0.17*	*0.03*	*0.13*	*5.56*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
0.50	*0.67*	*0.04*	*0.63*	*46.67*	*0.33*	*0.03*	*0.30*	*14.00*	*0.04*	*0.03*	*0.01*	*1.52*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
Forecaster # 65																				
0.2	*0.93*	*0.22*	*0.71*	*45.50*	*0.76*	*0.37*	*0.40*	*5.59*	*0.71*	*0.40*	*0.31*	*3.63*	*0.64*	*0.45*	*0.20*	*2.24*	*0.23*	*0.35*	*-0.12*	*0.55*
0.3	*0.79*	*0.20*	*0.59*	*14.67*	*0.53*	*0.18*	*0.35*	*4.99*	*0.41*	*0.28*	*0.13*	*1.76*	*0.29*	*0.23*	*0.06*	*1.35*	*0.15*	*0.20*	*-0.05*	*0.71*
0.4	*0.79*	*0.17*	*0.62*	*18.33*	*0.53*	*0.14*	*0.39*	*7.03*	*0.29*	*0.16*	*0.14*	*2.20*	*0.21*	*0.17*	*0.04*	*1.30*	*0.00*	*0.17*	*-0.17*	*0.00*
0.5	*0.71*	*0.16*	*0.56*	*13.57*	*0.47*	*0.09*	*0.38*	*8.78*	*0.29*	*0.10*	*0.19*	*3.66*	*0.07*	*0.12*	*-0.05*	*0.57*	*0.00*	*0.13*	*-0.13*	*0.00*
Forecaster # 84																				
0.2	*0.92*	*0.23*	*0.70*	*41.00*	*0.85*	*0.22*	*0.62*	*19.25*	*0.85*	*0.31*	*0.54*	*12.50*	*0.54*	*0.43*	*0.11*	*1.57*	*0.50*	*0.60*	*-0.10*	*0.68*
0.3	*0.85*	*0.11*	*0.73*	*43.08*	*0.69*	*0.13*	*0.56*	*15.11*	*0.77*	*0.13*	*0.64*	*22.38*	*0.08*	*0.17*	*-0.09*	*0.42*	*0.07*	*0.23*	*-0.16*	*0.26*
0.4	*0.77*	*0.08*	*0.68*	*35.93*	*0.62*	*0.09*	*0.52*	*15.68*	*0.46*	*0.06*	*0.40*	*12.37*	*0.08*	*0.08*	*-0.01*	*0.92*	*0.00*	*0.10*	*-0.10*	*0.00*
0.5	*0.77*	*0.07*	*0.70*	*47.14*	*0.54*	*0.07*	*0.46*	*14.58*	*0.31*	*0.06*	*0.25*	*7.56*	*0.00*	*0.04*	*-0.04*	*0.00*	*0.00*	*0.03*	*-0.03*	*0.00*

Note: w = decision threshold; H = hit rate; F = false alarm rate: PS = Peirce Score; OR = Odds Ratio.

Figure 1. Probability forecasts of quarterly real GDP declines at various horizons
(shades represent actual declines)

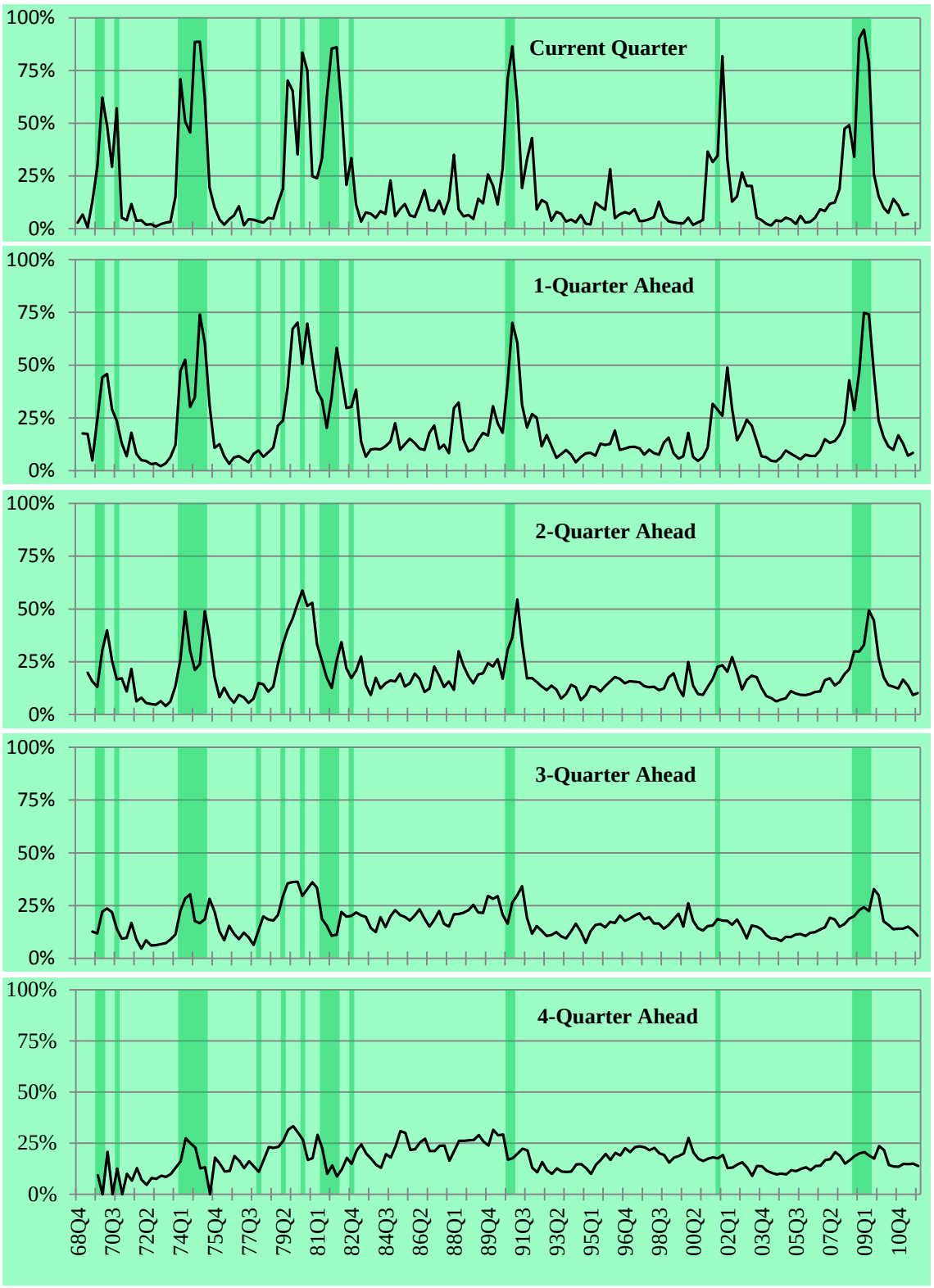
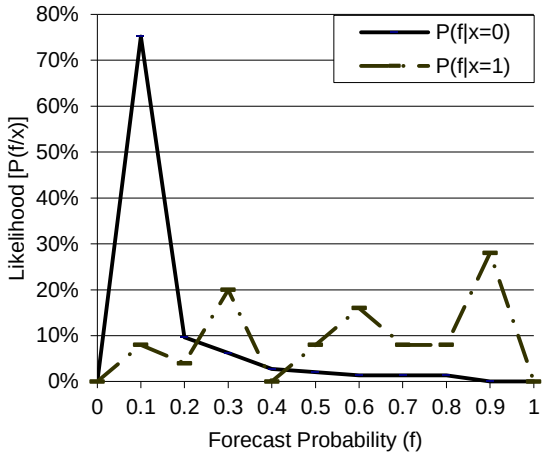
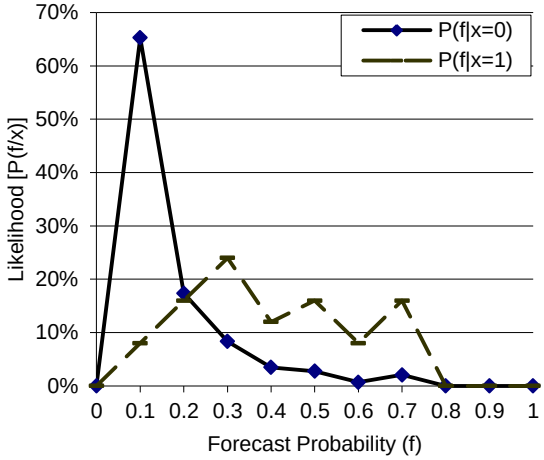


Figure 2: Conditional Likelihood, Q0 – Q4

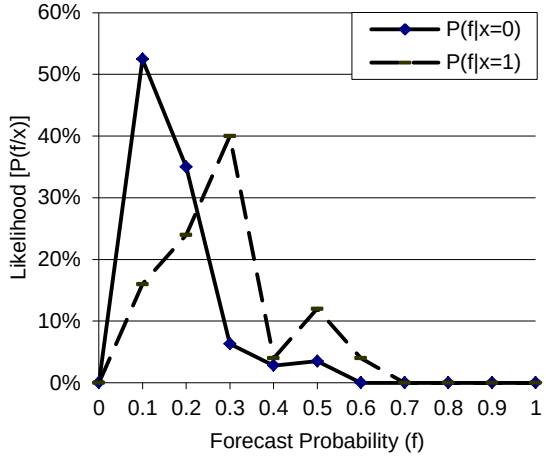
Quarter 0



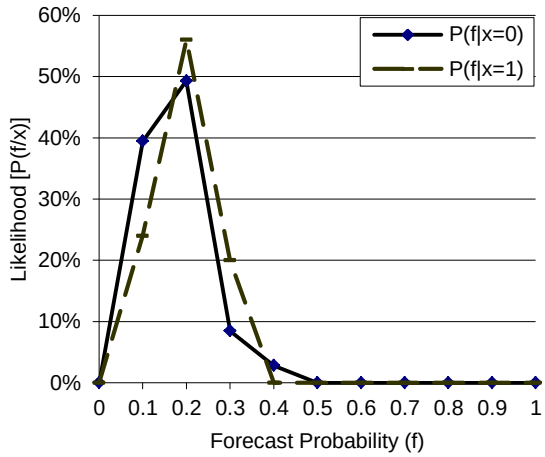
Quarter 1



Quarter 2



Quarter 3



Quarter 4

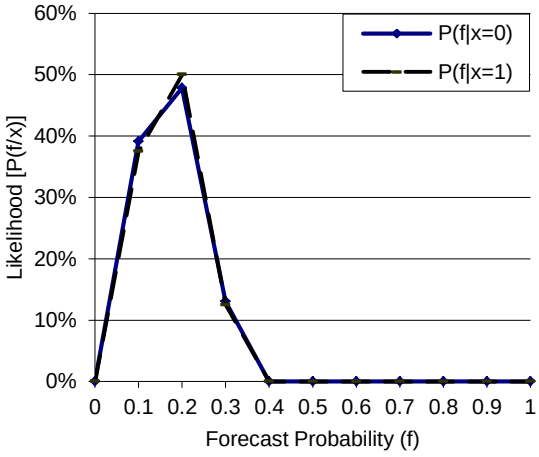
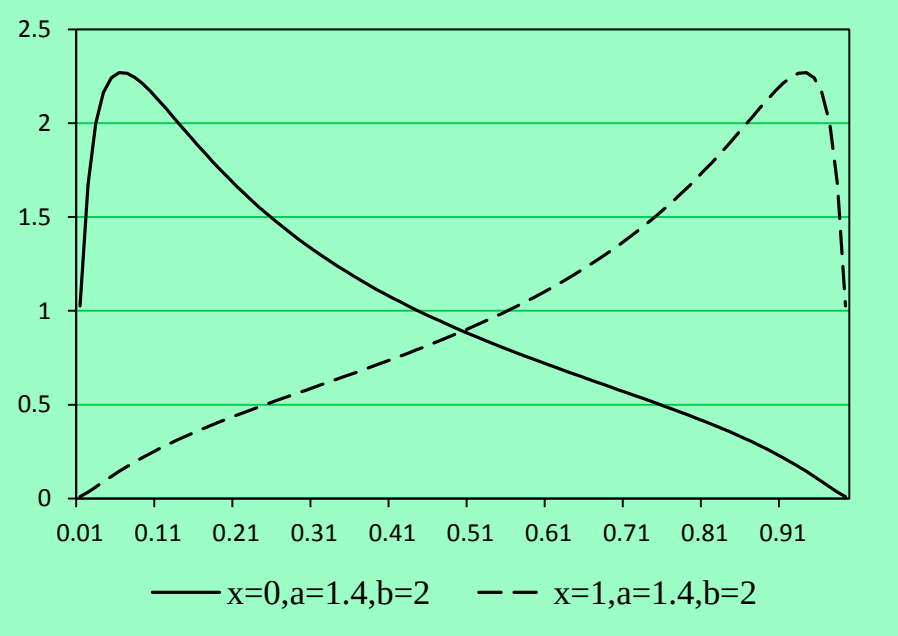


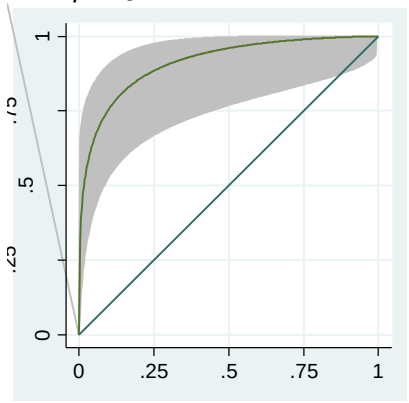
Figure 3: Notional Conditional Likelihooh Functions (Balanced Sample)



.

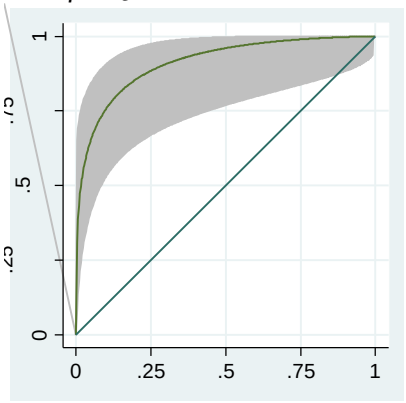
Figure 4: ROC Curves for Q0 – Q4 with 95% Confidence Band

ROC for Quarter 0 ± 95% Band



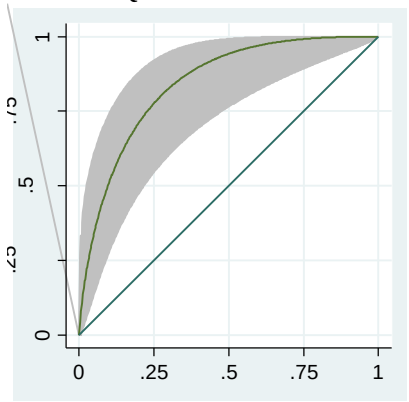
Y-axis: Hit Rate; X-axis: False Alarm Rate

ROC for Quarter 1 ± 95% Band



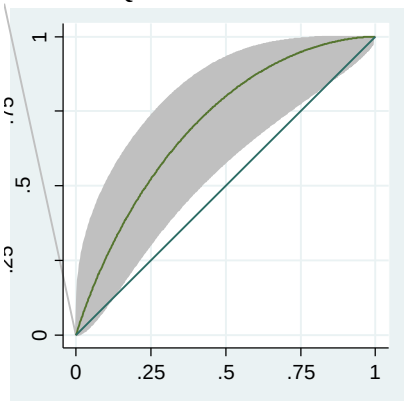
Y-axis: Hit Rate; X-axis: False Alarm Rate

ROC for Quarter 2 ± 95% Band



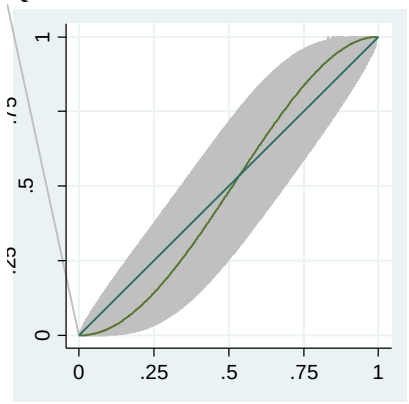
Y-axis: Hit Rate; X-axis: False Alarm Rate

ROC for Quarter 3 ± 95% Band



Y-axis: Hit Rate; X-axis: False Alarm Rate

Quarter 4 ± 95% Band



Y-axis: Hit Rate; X-axis: False Alarm Rate